Broadcast News is Good News

Francis Kubala
fkubala@bbn.com
BBN Technologies, GTE Corporation

Abstract

The past decade has been witness to the rapid maturation of automatic speech recognition technology. In the past nine years, the field has grown from small artificial research problems and discrete word commercial products and is now productively engaged on large real-world problems of unconstrained spoken language. In the research arena, the highly variable and content-rich domain of broadcast news has proved to be the right catalyst for producing robust scalable algorithms that methodically improve recognition accuracy every year. In the commercial arena, we've seen the appearance of wonderful shrink-wrapped dictation software and broad-based user acceptance. Now at the end of the millennium, speech recognition seems finally poised to realize its great promise. The Hub4 research effort of the 1990's and its recent focus on broadcast news have played a major role in placing automatic speech recognition on the launchpad.

Introduction

At the beginning of the decade, the DARPA-sponsored effort in Automatic Speech Recognition (ASR) took a huge step forward with the creation of the Wall Street Journal (WSJ) research test bed. The previous test bed, Resource Management (RM), was an artificially designed task whose language was defined by examples constructed by hand. The vocabulary was limited to 1000 words and the test data contained no words outside of the vocabulary. Speech was collected from subjects who read test utterances. Despite its considerable limitations, this hobbled test bed was profoundly successful in organizing a large research community around a common problem and in forging the tradition of periodic competitive evaluations on independent test data.

The advance to WSJ in 1990 solved the most grievous problems with the RM test bed. The underlying language for WSJ, although from written sources, was uncontrived. Huge corpora of text materials were gathered for language modeling. The vocabulary was initially constrained artificially but was gradually enlarged until the problem of inadequate word coverage with fixed system lexicons virtually disappeared around 64K words. Annual evaluations tracked steady improvements in recognition accuracy for 5 productive years. During this time, systems successfully scaled up to handle the large computational load of very large vocabulary problems.

With the introduction of the broadcast news test bed in 1995, the Hub4 research effort took another profound step forward. All of the deficiencies of the WSJ domain were resolved in the broadcast news domain. Most importantly, the fact that broadcast news is a real-world domain of obvious value has led to rapid technology transfer of ASR into other research areas and applications. On broadcast news, ASR has become an enabling technology.

This same decade has given birth to two other developments that will have a large impact on the deployment of speech recognition technology in the next millennium – fast cheap computing, and the Internet.

Commodity Computing

Advances in high-density microelectronics moved faster and farther than many of us imagined. In 1990, one of the main impediments to ASR research was computing. A Sun SparcStation2 with a 40MHz processor and 64MB of RAM sold for $20,000 (monitor not included)! At this workshop, we observe that a common platform used for the 10x real-time tests is the Intel 400MHz PentiumII with 512MB RAM. These machines can be purchased for under $2000. So in nine years, we've seen a 10x increase in desktop computer speed (roughly speaking), with a corresponding 10x reduction in price! That doesn't happen every decade.

For speech recognition, the timing could not have been better. As the restrictive compute ceiling lifted over the course of the decade, ASR system builders were able to explore more computationally expensive algorithms, perform more experiments, and work more efficiently overall by not having to be concerned with optimizing program size and speed at every turn. Moreover, the ever-expanding computational capacity made it possible to scale up quickly to large vocabularies and language models, to estimate model parameters from ever-larger training corpora, and to attempt big real-world problems like broadcast news.

The Ubiquitous Voice Channel

The other major development is, of course, the Internet, whose influence on ASR is only beginning. It is already difficult to recall the world of 1993, when the network consisted of *ftp*, *telnet*, and *finger*. Today, at the end of the millennium, the Internet continues to be mostly a *potentially* universal medium due the bandwidth limitations imposed by the network infrastructure. This restrictive environment will change soon, however, and it will change quickly.

Once again, the timing of the Internet growth could hardly have been better planned for ASR. While today's network bottlenecks work to clamp the explosive growth of the Internet, ASR applications, protocols, and standards are getting a chance to mature before they are broadly deployed. Voice over IP (VoIP) protocols have already been codified into world-wide standards (IETF's SIP, ITU's H.323) to support the emerging voice, data, and video services of the future. These integrated-media services provide a great opportunity for ASR applications since many of the difficult problems of interfacing ASR to the application and the communication channel, are solved by the protocol. In the same timeframe, speech and telephony APIs (SAPI, TAPI) have matured to help diverse voice applications interoperate. Dialog description languages like VoxML help to make human-machine dialog design for the Internet into a high-level scripting task. ASR will be a direct beneficiary of all of this infrastructure development.

Virtues of Broadcast News

The broadcast news domain has been an extraordinarily productive one for Automatic Speech Recognition (ASR) research. It has permitted research systems to demonstrate convincing annual advances on a problem of unquestioned value.

In 1995, when the broadcast news domain was introduced as the next research test bed to succeed WSJ, there was concern that the problem bar would be set too high, thereby inhibiting the steady advances in the field that had been taking place. In deference to this concern, the first broadcast news corpus was designed to reduce the impact of the great number of unknowns anticipated in the new domain. The training and test corpora were drawn from a single radio source, *NPR's Marketplace* -- a program of financial news. The training data was carefully marked by hand with a variety of features to denote the presence of corrupting music and noise, changes of speaking style and speaker, and changes in bandwidth. Furthermore, after the first evaluation on the Marketplace data, the initial problems encountered in segmenting the continuous input led in the following year to a return to discrete utterances by partitioning the test data into discrete speaker turns. Four years later, we see that none of these precautions was necessary.

Make the Problem Harder to Succeed

As we consider the progress in ASR accuracy that has been demonstrated annually on broadcast news since 1995, it's important to keep in perspective the wide variety of new challenges that have been overcome at the same time. Table 1 lists the major differences between the WSJ domain and its successor. It's clear that broadcast news introduces a bewildering variety of new problems for ASR.

The extreme variations in speaking style and accent as well as in channel and environment conditions are totally unconstrained. Since broadcast news is a highly edited composite product, no condition persists for very long. The sheer variety of conditions encountered guarantees that substantial gains in ASR accuracy could only be realized by robust and general solutions. Broadcast news is a superb *stress test* that requires new algorithms to work across widely varying conditions or to solve a specific problem without degrading any other condition. When considered in this light, the steady substantial progress we've observed in ASR accuracy on broadcast news over the last 4 years is truly remarkable.

WSJ	Broadcast News
Financial domain focus	National news focus
Written language domain	Spoken language domain
Simulated dictation	Real-world, found speech
Fixed train/test epoch	Train/test epochs advance
Same epoch for train/test	Training predates test
No story evolution	Natural story evolution
Data from live subjects	Unattended data capture
Discrete utterance	Continuous input
Single speaker session	Speaker number unknown
Single channel condition	Unconstrained conditions
Single speaking style	Every imaginable style
Native speakers	Non-native speakers
Good demonstrations	Great demonstrations
Speech only	Speech, video, text
Small research leverage	Huge leverage

Table 1. Comparison of WSJ and Broadcast News domains.

Table 1 also makes clear that the broadcast news domain is ideal for ASR R&D. There were immediate benefits to the program due to the greater efficiencies of data collection and corpus design. News is easy to collect and the supply of data is boundless. No live subjects or application simulations are needed. The data is *found speech*; it's completely uncontrived.

Real-World Benefits

News is a real-world domain of obvious commercial value that is familiar to everyone. It is rich in topical content that leads to very effective demonstrations. It is a multimedia source as well, including video and text in both on-screen and close captioning forms. There are many commercial uses of news, which lead to additional materials that can be leveraged in the R&D effort. For instance, several hundred million words of news transcriptions have been used in the Hub4 effort that were acquired through commercial channels.

R&D Leverage

Broadcast news focuses attention on temporal dimension of the problem: the language and lexicon evolve over time. News is filled with events, people, and organizations and all manner of relations among them. The great richness of material and the naturally evolving content in broadcast news has already leveraged its value into areas of research well beyond ASR. The Spoken Document Retrieval track of TREC and the Topic Detection and Tracking (TDT) program are supported by some of the same materials and systems that have been developed in the Hub4 arena. There are several other important areas in which ASR of broadcast news is currently being leveraged.

Name Extraction

Broadcast news is rich in named entities (people, places, organizations, etc) whose distribution changes naturally over time. It is an ideal domain for research in automatic name extraction. For the first time this year, name extraction has been incorporated as a research focus in the DARPA Hub4 evaluation. As a consequence, years of research in name extraction from text, organized around the recently discontinued Message Understanding Conferences (MUC), are now being leveraged anew on broadcast news speech. Moreover, the transition to speech has forced extraction technologies to deal with the loss of case, punctuation, and sentence breaks in the ASR transcript, and to accommodate the presence of recognition errors. In the process, automatic name extraction algorithms have become more general and more robust to variation in the source.

At the same time, the effort to extract names from speech has stimulated a new broad interest in dealing with the out-of-vocabulary (OOV) problem in ASR. Although builders of dictation systems have dealt with this problem, it has long been ignored in DARPA evaluations for a very good reason – errors due to OOV have always been a small minority of the total recognition error. Even with the totally unconstrained vocabulary of broadcast news, a 64K-word, fixed recognition vocabulary will cover new and temporally disjoint data with significantly less than 1% of the words missing. As long as reducing overall Word Error Rate (WER) was the objective for ASR, it made little sense to target OOV errors, since all errors were treated equally.

For the purpose of name extraction, however, we observe a much different picture. First of all, an OOV word in a name will almost always result in an extraction failure. Secondly, the impact of OOV for extraction is much greater than for ASR. In one large sample of broadcast news, nearly 4% of the words constituting the names were found to be OOV. Since names, on average, contain nearly 2 words each, the proportion of names containing an OOV word is about 7%. This accounts for a significant fraction of the total extraction error. The disproportionate impact of OOV on names will stimulate new work in rapid updating of recognition vocabularies and language models.

Toward Deeper Extraction

A new effort is underway to automatically extract information about events from news. Similar to recent work demonstrated in the last MUC evaluation, this new work attempts to push beyond the extraction of isolated named-entities to capture descriptors and attributes of the entities as well as relations among them. The early work in this area is scoped narrowly around a few event types such as natural disasters and terrorism and a small number of facts about them. Once a basic level of competence can be demonstrated in these first event types, scaling up is envisioned to hundreds of event types and thousands of facts.

Once again, broadcast news is an ideal domain for this advanced work. Since events persist over time, they will appear repeatedly in news from different perspectives and with different information as the event evolves. This property makes news an ideal source for event extraction. The first fruits of this pioneering work is likely to appear in next year's Hub4 evaluation, which has been dubbed *Event99.*

News on Demand Systems

The great richness, familiarity, and value of news media has spurred the creation of high-impact demonstrations of many of the advanced speech and language technologies developed with the sponsorship of DARPA. CMU's *Informedia*, MITRE's *Broadcast News Navigator*, and SRI's *Maestro* have all exploited the multi-media features of news producing a wide range of capabilities for browsing news archives interactively. BBN's *Rough'n'Ready* system concentrates on extracting as many structural features as possible from the audio alone.

These demonstration systems integrate various diverse speech and language technologies including ASR, speaker change detection, speaker ID, name extraction, topic classification and Information Retrieval. Advanced image processing capabilities are also demonstrated on the news video including scene change detection, face ID, and Optical Character Recognition from the video image. Demonstrations of these systems are very well received. Each of them is undergoing continuous and rapid development.

Speaker Change Detection

The unbroken nature of the broadcast news audio signal has led to a revival of interest in the problem of Speaker Change Detection (SCD). In broadcast news, the problem is complicated by the presence of music, noise, simultaneous talkers, the multiplicity of speaking styles and programming formats, and by the fact that the true number of speakers in any given episode is unknown.

The monolithic broadcast signal forced ASR systems to create novel means of segmenting the input, initially for the mundane reason of making the input manageable in size by splitting it into chunks without breaking words apart. But accurate location of the speaker boundaries is also important for ASR because of the many unsupervised normalization and adaptation techniques that are applied under the assumption that the adaptation data is produced by one speaker only. Because it is desirable to have as much speech as possible from any given speaker for unsupervised adaptation, ASR systems had to devise the means to cluster the segmented speaker turns into groups of segments originating from a common speaker. This also required a means for determining the number of speakers in the input. Nearly every system in this year's Hub4 evaluation has employed SCD and clustering to improve adaptation. Continued advances are expected in speaker change detection and speaker clustering in the service of ASR on broadcast news.

Hub4 98

Progress in ASR accuracy is graphically illustrated in Figure 1 for the last three years of Hub4 evaluations on the broadcast news domain. For each of the canonical signal/channel conditions defined for the Hub4 evaluations, the Word Error Rate is shown for the best system on that condition and the corresponding system combination results produced by the Rover system from NIST. Rover takes the output from multiple ASR systems and combines them into a single new recognition hypothesis by voting (with or without confidence weights) across the inputs.

Note that the three Hub4 test sets from 1996-1998 are not directly comparable overall because the different conditions are distributed in different proportions in each of the three test sets. Furthermore, there is not a big enough sample of the F5 Non-Native Speaker condition in any of the tests to include here. Also, there are no system combination results for the 1996 test. Finally, for this year's Hub4 test, we are showing two Rover results: one from the primary system results and one from the 10x real-time results.

The results from 1996 are the first ones achieved on a mix of typical radio and television broadcasts that are

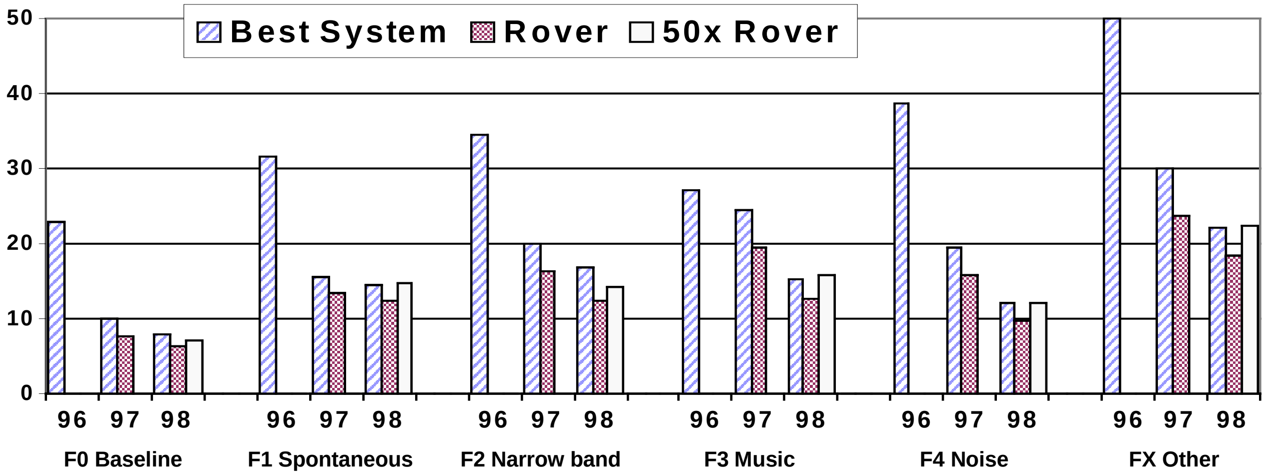

Figure 1. Comparison of Word Error Rate (WER) reductions for three years of Hub4 evaluations considering the best system per condition and system combination results from NIST's Rover.

presented to the ASR systems as a large monolithic waveform (unpartitioned input). The error rates for all conditions are elevated as is typical of initial trials on new problems. Very large WER reductions from the best system performance are observed in every condition except F3 (music) in the second year of the Hub4 news evaluation (1997). The modest F3 gain may be due to the better than expected first time performance. In the third year, 1998, there are significant gains in all conditions except F1 (spontaneous) which did surprisingly well in the second year.

These substantial and broad-based gains are due to many factors including increased quantities of training data and growing familiarity with the broadcast news problem. But, for the most part, the gains are due to many small contributions from new developments in modeling and unsupervised adaptation. Although none of the algorithmic improvements have single-handedly resulted in dramatic increases in accuracy, they have largely proved to be additive improvements and collectively have realized substantial gains over time. We have every reason to believe that these substantial annual gains can continue to be realized from numerous small additive improvements.

By 1998, WER for the baseline speech, which is the careful speaking style typical of anchor speakers, had returned to the levels (7-8%) achieved in the final years of the WSJ based tests. This parity was achieved despite the fact that the news speech had to be automatically extracted from the monolithic input whereas the WSJ input was presented to the system in grammatically complete discrete utterances. The 1998 results are also noteworthy for the robust performance demonstrated on the difficult and unpredictable F1-FX conditions. It is clear that ASR systems are becoming more robust.

System Combination and Simplicity

The Rover system combination results for 1997-98 show that in every condition, the performance of the best system can be improved by combining it with several inferior systems. The Rover improvements are consistently in the range of 10-20% relative to the best system. Note that in this year’s Hub4 test, most of the systems used a system combination strategy upon their own varied outputs. But system combination across systems has proved to be much more effective than upon several outputs generated by perturbations of a single system. This behavior indicates that a sizable portion of the remaining error can be modeled.

The 50x Rover results are produced by combining the top five performing systems in this year’s 10x real-time test. They are especially interesting because they equal or exceed the performance of the best system results, which were produced in several hundred times real-time typically. The 50x results illustrate that the extraordinary computational expenditures made in the primary tests are not producing corresponding value. It’s important to recall that each of the 10x systems used in the 50x system combination lost significant accuracy (typically 15-20%) by going to 10x real-time. The loss is due to simplification of the systems and reducing the search space with aggressive pruning. Still, the combination of 5 inferior systems equals or outperforms the best system for each condition, and it accomplishes this feat with a 4-8 times reduction in computation. One interpretation of this result is that most of the individual system complexity is not adding any value.

Operational ASR

With rapidly improving performance, convincing demonstrations on real-world data, and compelling content-based information management applications looming, it is an opportune time for ASR to move with purpose toward operational capabilities. These include real-time and low latency throughput, and most importantly perhaps, a return to simplicity in algorithm, operation, and implementation.

Low latency is important so that usable transcription output can be produced shortly after the data is captured, say within a few minutes. This implies that ASR needs to function in a near causal manner with only a short look-ahead. Typically, ASR systems represented in the Hub4 evaluations are deeply non-causal, involving many passes over the data in procedures that cannot be easily pipelined. When the data is an hour long, it still takes an hour to get usable output from a non-causal real-time system.

It is also important to recover as much simplicity as possible in our ASR systems so that they can be deployed widely and reliably. As the 50x Rover results have shown so clearly, much of the recent complexity of large ASR systems is unnecessary. We can expect a good deal of application pull toward system simplification in the near future as the transfer of our successful ASR technology continues apace.